DOCUMENT RESUME

ED 310 161 TM 013 820

AUTHOR Yamamoto, Kentaro
TITLE Hybrid Model of IRT and Latent Class Models.
SPONS AGENCY Office of Naval Research, Arlington, Va.
PUB DATE Jan 89
GRANT N00014-82-K-0604-NR-150-495
NOTE 64p.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Statistical Data (110)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Ability Identification; Achievement Tests; *Cognitive Structures; Diagnostic Tests; Estimation (Mathematics); *Latent Trait Theory; Mastery Tests; *Mathematical Models; Maximum Likelihood Statistics
IDENTIFIERS EM Algorithm; *Latent Class Models; Parametric Analysis

ABSTRACT

This study developed a hybrid of item response theory (IRT) models and latent class models, which combined the strengths of each type of model. The primary motivation for developing the new model is to describe characteristics of examinees' knowledge at the time of the examination. Hence, the application of the model lies mainly in so-called achievement, diagnostic, and mastery testing environments. The treatment of multidimensionality is central to the hybrid model. The new model's essential characteristics involve its abilities to represent a certain cognitive structure, to incorporate test users' informal task and error analysis as well as formal exhaustive analyses, and to relate to IRT in order to facilitate assessment of inter-item equivalency and performance of differential item analysis. Parameter estimation is central to the attempt to develop the hybrid model. An EM algorithm is used to compute maximum-likelihood estimates from incomplete data. Two experiments were undertaken to assess the model. Overall, results indicate that the hybrid model provides unusual freedom based on its unification of continuous and discrete item response models. By varying group and state probabilities, the model can range from one extreme of IRT-only to the other extreme of latent-classes-only models; any combination of these two models can also be arranged. These latent classes do not have to be ordered in any manner; hence, they may be able to accommodate individual differences of the order of learning materials. A 56-item list of references and 12 data tables are included. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED310161

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
KENTARO YAMAMOTO
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

HYBRID MODEL OF IRT AND LATENT CLASS MODELS

by

Kentaro Yamamoto

Educational Testing Service
Princeton, New Jersey

January, 1989

This research was partially supported by Grant No. N00014-82-K-0604, NR 150-495 of the Office of Naval Research awarded to Kikumi Tatsuoka. I am grateful to Kikumi Tatsoka, Maurice Tatsuoka, Robert Linn and William Stout who were particularly helpful during the time I was writing my dissertation on which this paper is based. I also appreciate Murray Aitkin and Robert Mislevy for their comments and discussions with me at the later stage.

TM013820

BEST COPY AVAILABLE

INTRODUCTION

Concern has been increasing recently over the inability of Item Response Theory (IRT) to express cognitive structure, such as whether a response was produced by a rule of operations or a particular misunderstanding, or to give descriptive information of knowledge possessed by examinees. This lack of capability has spawned several new developments in IRT, and in latent class models as well.

Since the pioneering work in latent structure analysis of Lazarsfeld (1960), the gap between the continuous model and the discrete model of latent structure advocates has become wider and wider. This departure is due to the differences in the basic understanding of the structure of human abilities and how we intend to use the results from applications of these models.

IRT is used mainly to aid the development of large-scale tests, to evaluate tests and test scores, to study item bias, to create computerized adaptive tests and to equate the multiple forms of tests. The most practical use is to attach scale values to examinees. But IRT is incapable of saying anything specific about the cognitive structure of a given examinee. Let us define cognitive structure as a structure of knowledge or understanding of a given cognitive domain, such as arithmetic, electrical circuits and so on.

The latent class models, on the other hand, are used for assessing the examinees' traits or acquisition of competency. These models are used to make statements about the examinee's cognitive characteristics. These "characteristics of examinees" may be cognitive structures of understanding or misunderstanding.

Many researchers of the cognitive science in the past expressed cognitive structure by a set of response vectors (Birenbaum & Tatsuoka, 1981; Brown & Burton, 1978; Riley & Greeno, 1988; Siegler, 1981). Here, cognitive structure is inferred from the response vectors through inductive reasoning. Therefore, one or a set of response patterns is associated with a given cognitive structure. The relationship between items and cognitive structure relies heavily on task analysis. I will utilize this method to represent a structure of cognition by a particular response vector or a set of them.

Assumptions and usage of the two models seem to remain incompatible, except for the assumption of conditional independence of responses to different items given the value of a parameter which defines an examinee's ability. Yet interest is increasing in a model that can represent some form of cognitive structure and also the relationship among items, test forms, and examinees. That is, a model is needed which retains the benefit of IRT and yet possesses the specificity of the cognitive structure of the latent class model. Such a model could be used for testing educational achievement.

Unidimensionality, the key assumption of IRT, poses a problem in fitting data also. The evidence is overwhelming that even a restricted subject area, such as arithmetic, is apparently multidimensional (Birenbaum and Tatsuoka, 1981; Tatsuoka and Tatsuoka, 1982; Lord, 1980). There are several ways to cope with the multidimensionality of data sets, and the choice is determined by the understanding of the cause of this multidimensionality itself.

There are several approaches to handling the multidimensionality in items. The most common one is to assume that the items measure multiple distinct traits (Lord, 1980; McDonald, 1981). This view encourages the development of alternative multidimensional IRT models (Samejima, 1969, 1972, 1974; Sympson,

1977; Reckase, 1982, 1985), to examine the robustness of unidimensional IRT to multidimensional data (Wang, 1986, 1987), to find a way to separate items to sets of acceptable unidimensional items (Bejar, 1980), and to develop statistical methods to determine the degree to which the data deviate from unidimensionality (Holland, 1981; Stout, 1984, 1986).

Another approach is to treat dimensionality as a result of the interaction between the items and the examinees under specific conditions (Bejar, 1983), and these conditions may be within the examinees themselves or the environmental constraints (Mislevy, 1983, 1986; Tatsuoka, 1983). This is based on the notion that the testing conditions such as speededness or other constraints cause multidimensional items (Traub, 1983). Also erroneous understanding which is "hopefully" a transient state of understanding of the material can cause the multidimensionality (Tatsuoka, 1984, 1985).

In his review of the state of latent trait measurement theory, Traub (1983) concluded that educational achievement tests do not fit unidimensional item response models. The multidimensionality of items now seems to be an unavoidable fact, and test theory has to take into account this fact in order to derive useful information from testing.

Considering the inability of IRT to represent cognitive structure, much less the change in it, and the inherent ambiguity of "ability", the choice of a multidimensional ability space to describe a data which deviate from unidimensionality of items does not seem to be promising. Especially, compensatory multidimensional IRT does not apply to many of the tests we have. For example, word problems in mathematics cannot be answered correctly if one lacks either computing or reading ability. The other reason for the

inappropriateness of using continuous model alone to express learning is the existence of the so called "ah-ha" effect of learning. The notion of single-event learning cannot be incorporated easily into a purely continuous model.

Task analysis is an extremely time consuming activity, and hence it is not routinely done prior to test administration, even though it can provide extremely useful information about the construction of items. The fact that no currently available psychometric model can incorporate the result of task analysis at the time of parameter estimation is also a disincentive. Consequently items may be chosen to be included in an operational test based on the degree to which they fit the model, instead of the intended measuring objectives (i.e., construct validity). This is to say that in the past, item construction and inclusion into a test battery reflected the idea of content validity. "The content validity is essentially ad hoc" (Loevinger, 1957), and current researchers in the validity of tests agree (Angoff, 1986; Messick, 1980). Consequently, to talk about validity one has to focus on the psychological construct which the items presumably measure, and validity is the degree to which the data from the test relate to the underlying construct (Messick, 1984). Hence, construct validation includes not only quantitative data analysis but also logical task analysis, as well as development of a measurable construct itself. In IRT construct validity is not a part of parameterization; i.e., there is no place for a psychological construct other than the so-called "ability", theta. The definition of this theta still remains uncertain in the field of psychometrics, along with its dimensionality.

Even in the absence of extensive item analysis, experience of many years of teaching and in-class testing by teachers may be thought of as informal error analysis, and the author believes that this is one of the greatest assets to the testing environment. It is currently not used at all in many phases of test development. This is especially true at the time of model fitting (parameter estimation of the response data.)

The characteristics that the new model should have are; a) ability to represent a certain cognitive structure of examinees, b) ability to incorporate test users' informal task and error analysis as well as formal exhaustive analyses, c) ability to relate to IRT so that inter-item equivalency can be assessed and differential item analysis can be performed.

The primary motivation to develop the new model is to describe characteristics of examinee's knowledge at the time of examination. Hence the application of the model is mainly in so-called achievement testing, diagnostic testing, and mastery testing environment. For many examinees, the regularities in response patterns might allow IRT to provide a reasonable summary; while for others, with characteristic alternative patterns, the distinct "non-scalable" classes wold be appropriate to summarize performance. The hybrid model described here offers the capability of modeling such a mixture. It is also possible that in some "ability tests", some examinees may be characterized by particular response patterns of subset of items. Then the new model would be applied successfully to those tests as well.

PREVIOUS MEASUREMENT MODELS

ITEM RESPONSE THEORY

IRT provides a mathematically explicit relationship between unobservable hypothesized ability and observed performance on a test item. Like most of item response structure models, both logistic and normal ogive item response models, such as Rasch's (1960) Logistic model and the two and three parameter models, (Birnbaum, 1968; Lord, 1980) assume that an examinee's probability of answering an item correctly increases as the ability value increases. This assumption is called monotonicity. The distribution functions that traces logistic and normal ogives are quite similar; the reason for the popularity of the former over the latter is its mathematical convenience. The logistic distribution has a closed form and can be evaluated easily.

The number of parameters employed to specify the function distinguishes the various item response models. The Rasch model uses one parameter, the item difficulty, which is often expressed by b. It specifies the abscissa of the point of inflection, which is (b,.5) for one and two parameter models. The probability of a correct response on item i for an examinee with ability theta can be expressed using one parameter item response model as

$$(2.1) \qquad P_i(\theta) = \frac{1}{1 + \exp[-D(\theta - b_i)]} \qquad \text{(Logistic)}$$

$$= \Phi(\theta - b_i) \qquad \text{(Normal Ogive)}$$

where D is a constant used to make the logistic and normal ogive curves nearly equivalent and often set to 1.7.

The two parameter item response model can be obtained by using a discrimination parameter a_i (related to the slope of the item response function at $\theta = b_i$) in addition to difficulty parameter. Most frequently, two parameter IRT model is expressed as

$$(2.2) \qquad P_i(\theta) = \frac{1}{1 + \exp[-Da_i(\theta - b_i)]} \qquad \text{(Logistic)}$$

$$= \Phi[a_i(\theta - b_i)] \qquad \text{(Normal Ogive)}$$

There are advantages and disadvantages for using each model with a certain number of parameters. But unsuccessful attempts to fit the one-parameter model to the data strongly indicates the necessity of having varying discriminating parameters (Lord and Novick, 1968, p402; Bock, 1967). Recent developments in item parameter estimation procedures have made it possible to estimate item parameters of IRT model with more than one parameter accurately (Lord, 1974; Bock & Aitkin, 1981).

The 3-parameter model introduced by Lord (1952) and Birnbaum (1968) has more apparent "accuracy" in representing the multiple choice item administration condition by having a "guessing" parameter c_i. The conditional probability of a correct response given θ for the 3-parameter model is a sum of c_i and the product of $(1-c_i)$ and $P(\theta)$ of the 2-parameter model. But it lacks a sufficient statistic for the ability parameter, requires a large number of observations, and the estimation of the c parameters is unstable when there are few respondents with low abilities. If one intends to use the item response model in a relatively small sample setting, then a smaller

number of parameters to be estimated is desirable. In the present study, the 2-parameter IRT model will be used. It has the advantage of having sufficient statistics (Birnbaum, 1968), and the estimation is more stable with a smaller sample size. Extending the algorithm developed using 2-parameter model to the 3-parameter case is straight forward.

In sum, IRT has proven to have great utility in the study of bias, equating of multiple forms, and scaling of achievement of examinees. Yet it has limitations, such as the cognitive structure of examinees and the prior error analysis of items, these cannot be incorporated in any way to the model, hence the parameters of the model cannot reflect the accumulated experience on the items. Concerning an examinee's ability, IRT provides only a scale value which is not specific and can be used only to place the examinee on a unidimensional continuum of overall proficiency.

Latent Class Models

In the traditional line of psychometrics, the primary purpose of measurement is to attach some scale values to each individual. Of course, many things could be said about the population of examinees and items being used, but as far as the individual goes, ordering is the only result we get and use. Consequently a continuous latent trait is commonly assumed. The use of the results as feedback to change examinees is not a main concern, and possibility of feedback resulting in discontinuous change in ability has not been seriously treated. Even though the assumption of continuous latent ability works well in the population, the direct application of continuous latent ability to an individual poses a serious problem. It is not very

likely that as one increases in latent ability θ, the probability of answering correctly on a given item i, $P_i(\theta)$, increases continuously. A rather more realistic trace would exhibit discrete levels of probability to reflect the occurrence of learning. After all, single event learning cases can not be expressed with a continuous model. IRT can model successfully discrete levels of probabilities in terms of the population in the special case that the number of levels is two and "ordered". Ordered levels means that item correct-response probabilities of two levels, level 1 and level 2, have inequality for all items i, and $P(x_i=1|\text{level 1}) < P(x_i|\text{level 2})$. In other words, the propensity of correct response is greater among subjects who belong to level 2 than those in level 1 for all items. This adequacy of IRT for modeling two discrete levels of learning soon diminishes when levels are more than two and are not ordered; that is, when the previous inequality holds only for the part of items. Even when multiple levels are ordered completely, the shape of the ICC of IRT alone can be a cause of misfit. Therefore when we shift the focus of our study from population latent structure to individual latent structure, the appropriate model has to be used.

There are two ways to cope with discontinuous $P_i(\theta)$. One is to view theta as being continuous, but there exist some disjoint sets of examinees whose response patterns are not explained by a θ-based model. The other is the view that theta is discrete and may not be even ordered; hence $P_i(\theta)$ is also discrete. Tatsuoka's rule space uses the first approach, and latent class models use the latter approach.

Since Lazersfeld introduced latent structure analysis, several improvements have been made. I'll describe characteristics of latent class models in general. The mathematical aspect of the models are mostly

consistent throughout the works of Lazersfeld and Henry (1968), Goodman (1974), Dayton and Macready (1976), Wilcox (1979), and Paulson (1982).

The vectors of responses to items by subjects make up the data. The responses are not necessarily dichotomous for the model to work but for now only dichotomous item responses are treated. Latent classes are mutually exclusive and together exhaustive, and finitely many latent classes exist. Every examinee belongs to one of the classes. The propensity of a correct response to each item i from any subject who belongs to a given class is called the conditional probability of correct response to that item given the class membership, e.g. $P(x_i=1|class1)$. The key assumption that is equivalent to the local independence of IRT is the conditional independence given the class of the subjects, i.e. $P(x_i=1|class1,x_{i'}=1) = P(x_i=1|class1,x_{i'}=0)$. The probabilities of class memberships are given by P(k). The model is completely described by the matrix of conditional probabilities and distribution of classes.

Table 1 shows an application of the constraints of Paulson's model (1986a, 1986b), which is a general unordered latent class model for signed number arithmetic. The letters L and S of item type represent larger or smaller of absolute values of 2 numbers. For example an item type (-L+S) means that add a positive number smaller in absolute value to a negative number larger in absolute value, for example (-5+3). The e's are execution errors associated with each rule. Let us consider separating responses into 2 components, absolute value and sign components, hence scoring can be carried out independently as well. The table shows the absolute value component. As an example, consider the "always add" rule. Except for one item type (-L+-S), faithful execution of this erroneous rule produces wrong answers on all other

item types. Another additional constraint was placed here. Consider two item types (-L+S) and (-L+-S), execution of the erroneous "always add" rule produces 0 and 1 respectively. The constraint is that the execution error, given a latent class, is constant for all items, hence an execution error response on (-L+S) is actually a correct response and its conditional probability of a correct response is ϵ_2. An execution error on (-L+-S) is a wrong response, hence its probability of correct response is $1-\epsilon_2$.

<insert table 1 about here>

At this point it should be clear that once these conditional probabilities P_{jk}'s and probabilities of class membership P(k)'s are estimated, latent class models become ordinary compound binomial models. Then we can have values of accuracy and consistency of classification, and the fit of the model can be evaluated by covariance analysis (Yamamoto, 1983). Most importantly, "qualitative" information on the states of the examinees' knowledge can be assessed.

Even though the latent class model provides different information from IRT, the difficulty of determining the number of classes, and the mutual exclusiveness of latent classes poses problems. It is a very difficult task to determine the number of latent classes especially when they are ordered. Haertel (1984) found that the latent class model with two classes fits reasonably well to an achievement test that is customarily considered to be a more appropriate test for IRT.

RULE SPACE

The rule space model was developed by K. Tatsuoka and her coworkers to solve a specific classification problem, that is to classify the response patterns to the so called rule of operations that one can use in solving a problem in some well defined domain such as arithmetic, or algebra (Birenbaum & Tatsuoka, 1981; Brown & Burton, 1978; Riley & Greeno 1988). Naturally, a correct rule produces the right answer to all the items, but sometimes a wrong rule (erroneous rule) may produce the right answer to a subset of items. Rule space starts with the mapping of all response patterns into a set of ordered pairs (θ,ζ), where θ is the estimated ability value of 1 or 2 parameter IRT and ζ (zeta) is an IRT-based caution index. The scale value of ζ for a response vector is an inner product of 2 residual vectors, namely $(\underline{P}-\underline{X})*(\underline{P}-\underline{T})$ where $\underline{P}$ is the conditional probability of IRT at the estimated theta of response pattern $\underline{X}$, and $\underline{T}$ is the mean of $\underline{P}$ so it is a constant. Zeta indicates how well respondents' patterns accord with the model; low values for good fit, and high values for bad fit.

Mapping all response patterns into the theta by ζ space produces a distribution of points in the space. Tatsuoka (1983) showed that a response pattern and a point in the space has a virtually one to one relationship, "similar" response patterns have similar values of theta and ζ, hence they lie in proximity. After all the response patterns including those from the erroneous rules are mapped on the rule space, one would find clusters of points, each cluster representing a unique erroneous rule. It follows that if two erroneous rules are conceptually similar then the similarity exists also in the response patterns produced by the perfect application of the rules then

two clusters would be close or might even overlap with each other. Within a cluster, the distribution of the points is called bug distribution (Tatsuoka & Tatsuoka, 1987), and is used to classify a response vector and diagnose an examinee's cognitive error that results in his or her response pattern.

Rule space is not without problems. In particular, the use of IRT-based $P_i(\theta)$'s for a latent class with a response pattern which differs greatly from a vector of class-based $P_i(\theta)$'s for conditional probabilities given the class is troublesome. The reason is quite simple; let us suppose that we have two erroneous rules of interest, each rule having a unique response pattern associated with it. Suppose further that these two response patterns happen to yield identical estimated theta values, or at least one that are so close that the difference is negligible. Then regardless of class membership IRT-based conditional probabilities of correct response on item remain the same, that is $P(x_i=1|\theta,\text{class I}) = P(x_i=1|\theta,\text{class II}) = P(x_i=1|\theta)$ for all item i. Since two classes with the same theta value have identical conditional probabilities, the likelihood to produce a response pattern given a class is also identical for both classes. This leads to some very unlikely results.

For example, let us assume there are two classes I and II, each specified by response vectors 11000 and 00111 respectively. Also both would have identical estimated theta values when a's are 2.0, 1.0, 1.0, 1.0, and 1.0 respectively and by supposing that conditional probabilities are .3, .3, .5, .7, and .9 because of certain b values. Then the slip probabilities for class I, ie. propensity to deviate from the class I response vector, are .7, .7, .5, .3, and .1. Likewise for the class II, the slip probabilities are .3, .3, .5, .3, and .1. The class II seems very reasonable, the probability to deviate

from the response pattern given the subjects in the class II are less than or equal to .50. But with class I, for the first two items, slip probabilities are greater than the probabilities to produce responses of the class I response pattern. This is to say that subjects who belong to class I and are identified by the class I response pattern 11000 are modeled as having greater probability to deviate from 11000 response pattern than to conform to it. Also among subjects who belong to class I have actually greater probability to produce 00111 response pattern than their own 11000 pattern. This, and an observation of distribution of response patterns around an erroneous rule which has a high zeta value in the rule space of arithmetic data, suggest that some modification of conditional probabilities for rules which have high absolute values of zeta may be necessary.

A correction of conditional probabilities was proposed by Yamamoto in 1985, who used linear equation to combine a class response pattern and conditional probabilities of IRT's $\underline{P}(\theta)$. This produces corrected conditional probabilities which lie between class response pattern and $\underline{P}(\theta)$ of IRT.

Characteristics of a New Model

Now a general framework of a new model that follows a traditional probabilistic modeling approach can be set up to combine IRT and latent class models. From the standpoint of IRT, added latent classes represent systematic deviations from IRT; and also the classes are additive. Thus the population is a mixture of respondents who are well modeled by IRT and classes of respondents who are not. For example, a set of examinees may contain some who randomly guess at all items, and hence do not fit IRT, along with some

examinees who really try to answer correctly; their responses fit IRT. This situation can be represented correctly by having random guessing class added to the IRT model. For those latent class model advocates, the IRT part of the hybrid model can be thought of as representing the classes which are ordered; hence the ordering of examinees is the most prominent information about the nature of their abilities. This would provide a greater flexibility on the usage of partial knowledge about the cognitive structure of the content that we intend to measure. Also in the rule space environment, rules with high zeta values would have conditional probabilities significantly different from those of IRT; the subjects using such a rule can be thought of as latent classes in the hybrid model.

HYBRID MODEL

In order to understand a model, a concrete example is always helpful. Imagine that there is a data set which contains dichotomous responses, correct or incorrect, on items by each examinee. The population from which the examinees were randomly selected has two distinct groups of people; one group responds to items in accordance with IRT and the responses of the other group are associated with unique understandings or misunderstandings of the material being tested. They will be called IRT and Latent-Class groups, respectively, hereafter. It should be noted that there may be several latent classes in the Latent-Class group, each representing a unique understanding and misunderstanding. Each such latent class is characterized by a specific response pattern (i.e. an idealized response pattern) and the propensity of correct responses of an examinee from this group resembles this idealized response pattern. The actual propensity of correct response to an item would

differ from the 1.0 or 0.0 of the idealized response. The probabilities of execution errors will appear as parameters in the hybrid model.

If one is familiar with Tatsuoka's (1983) "Rule Space", or Paulson's (1982) "Latent State Model", each latent class can be thought of as a specific erroneous-rule class. Each latent class is characterized by a unique misunderstanding of operations and hence the consistent application of this misunderstanding produces an unique response pattern. The deviation of response patterns of examinees in the latent class group is caused by execution error or application error of the erroneous rule. In short, there are two groups, IRT and Latent-Class, and each latent class within the latent-class group is represented by a specific idealized response pattern.

The model starts with the following assumptions.

1) Each examinee belongs to IRT group or one of the classes within the latent class group. The classes in latent class group and IRT group are mutually exclusive and together exhaustive.

2) Class and group membership are unknown; hence the data cannot be separated into multiple subsets of data prior to parameter estimation.

3) Within the IRT group, both of the standard assumptions of IRT apply: a) local independence, b) unidimensionality.

4) Responses are conditionally independent given a latent class; this is equivalent to the local independence of IRT.

5) Classes are to be expressed by specific "ideal" response vectors.

In addition to the assumptions above, a fact about the mixture distribution needs to be noted: that is, the ability distribution in the IRT group is independent of the mixture proportions of IRT and latent-class groups.

Parameters of the Model

In addition to the usual IRT item parameters and examinees' ability parameters of the specified IRT models, there are conditional probabilities of correct response on each item given a latent class, and the proportion of examinees in each class. The model is a mixture of both IRT and Latent Class models, and a parameter for the proportion of the examinees in the two groups specifies the mixture.

The notation is as follow:

X_{ij} - 1 if subject j is correct on item i

0 if subject j is incorrect on item i

I - number of items i=1,...,I

J - number of subjects j=1,...,J

K - number of groups (including the IRT group and all classes in the latent-class group)

$P(\gamma=k)$ - probability of a randomly sampled subject being in a class k

k = 1 for IRT

k =2,3,...,K for latent classes

ζ_i - item parameters (a_i,b_i), and latent class parameters

θ_j - ability parameter of subject j

$f(\theta)$ - probability density function of θ in the IRT group

Γ_{jk} - class membership matrix; for given subject j, only the Γ_{jk} with k standing for the class to which he belongs is 1 and all other $\Gamma_{jk'}$ are 0

Q_t - number of discrete values of theta within the IRT group, used to approximate the integration over θ by the summation across discrete θ values, $q=1,\ldots,Q_t$

$A(\theta_q)$ - quadrature weight associated with θ_q (IRT)

R_{1iq} - number of correct responses on item i among IRT subjects who are at a fixed theta θ_q

N_{iq} - number of attempts on item i among IRT subjects who are at a fixed θ_q

R_{1ki} - number of correct responses on item i among subjects in a latent class k

N_{ki} - number of attempts on item i among subjects in a latent class k

Logistic IRT (2-parameter)

$$(3.1)\quad P_{1ij} = P(x_{ij}=1|\theta_j,\underline{\zeta}_i) = \frac{1}{1.0+\exp(-Da_i(\theta_j-b_i))}$$

$$(3.2)\quad Q_{1ij} = Q(x_{ij}=0|\theta_j,\underline{\zeta}_i) = 1.0-P_{1ij}$$

Latent Class

$$P_{ki} = P(x_i=1|\ \Gamma_{.k}=1)$$

$$Q_{ki} = P(x_i=0|\ \Gamma_{.k}=1) = 1.0-P_{ki}$$

By the assumption of conditional independence in the IRT model the conditional probability of observing a response vector $\underline{x}_j$ for a subject j from the IRT population is given by

$$P(\underline{x}_j|\theta_j,\underline{\zeta},\Gamma_{j1}=1)=\prod_i^I P_{1i}(\theta_j,\underline{\zeta}_i)^{x_{ij}}\ Q_{1i}(\theta_j,\underline{\zeta}_i)^{1-x_{ij}}.$$

If a subject j is in the latent class group k ($k\geq 2$) then the conditional probability is given by

$$P(\underline{x}_j|\Gamma_{jk}=1) = \prod_i^I P_{kij}^{\ x_{ij}}\ Q_{kij}^{\ 1-x_{ij}}$$

By conditional independence in the latent class model, for all j with $\Gamma_{jk}=1$, P_{kij} and Q_{kij} remain constant over j. That is to say, all subjects who belong

to a latent class have the same conditional probability of responses given an item.

If subjects are selected from the mixture population described earlier, the conditional probability of response matrix given ability, item parameters and $\underline{\Gamma}$ is given by

$$(3.3) \qquad P(\underline{x}_j|\theta_j,\underline{\zeta},\Gamma)= \begin{cases} \prod_{i=1}^{I} P_{1i}(\theta_j,\underline{\zeta}_i)^{x_{ij}}\, Q_{1i}(\theta_j,\underline{\zeta}_i)^{1-x_{ij}} & \text{if}(\Gamma_{j1}=1) \\ \prod_{i=1}^{I} P_{kij}{}^{x_{ij}}\, Q_{kij}{}^{1-x_{ij}} & \text{if}(\Gamma_{jk}=1) \end{cases}$$

The above expression can be thought of as $L(\underline{\theta},\underline{\zeta},\underline{\Gamma}|\underline{X})$, once we have responses $\underline{X}$.

In order to utilize this model, the number of latent classes, the idealized response pattern of each class, and the constraints of latent class parameters must be specified. Then the hybrid model algorithm estimates the following: IRT parameters for each item and each examinee, a vector of conditional probabilities of each latent class, and distribution of examinees across the IRT group and latent classes. Also since the posterior distribution is readily available classification of examinee to one of the categories (IRT and latent classes) can be carried out, for example, using the likelihood ratio test, if decision criteria were provided. It also should be noted that the measurement problem that the rule space model tries to solve is the same problem that the hybrid model is developed to solve.

METHOD TO ESTIMATE THE PARAMETERS OF HYBRID MODEL

Marginal Likelihood Function

The conditional probability of observing a response vector $\underline{x}_j$ given θ, and ζ is given by

$$(4.1) \qquad P(\underline{x}_j|\theta,\zeta) = \prod_{i=1}^{I} P_{1i}(\theta)^{x_{ij}} Q_{1i}(\theta)^{1-x_{ij}}$$

The probability of observing the response pattern $\underline{x}_j$ of a randomly sampled subject from a population with a density function $f(\theta)$ conditional on known item parameters, can be obtained by the integral of the above expression over the IRT population, i.e.

$$(4.2) \qquad P(\underline{x}_j|\zeta) = \int_\theta P(\underline{x}_j|\theta,\zeta)f(\theta)d\theta$$

This is the marginal probability of response pattern $\underline{x}_j$ given ζ, and $f(\theta)$. Let us consider the mixture population, where Γ has a density function $p(\gamma)$, i.e. the probability density of classes, since we can use the density function in place of Gamma if we randomly sample subjects. Ordinarily, $p(\gamma=1)$ is the proportion of subjects who are in the IRT group, and $p(\gamma=k)$ where $k\geq 2$ is the proportion of subjects who are in the latent class k. Rewriting $P(\underline{x}_j|\zeta)$ as $P(\underline{x}_j|\zeta,\gamma=1)$ by letting ζ include within latent-class parameters as well as item parameters of IRT group, then the conditional

probability of $\underline{x}_j$ given class parameters is

$$P(\underline{x}_j|\underline{\zeta}) = \sum_{k=1}^{K} P(\underline{x}_j|\underline{\zeta},\gamma=k)\cdot P(\gamma=k)$$

$$= L(\underline{\zeta}|\underline{x}_j)$$

Hence

$$(4.3) \qquad L(\underline{\zeta}|\underline{X}) = \prod_{j}^{J} L(\underline{\zeta}|\underline{x}_j)$$

$$= \prod_{j}^{J} \sum_{k=1}^{K} P(\underline{x}_j|\underline{\zeta},\gamma=k)\cdot P(\gamma=k)$$

The above expression can be thought of as a likelihood function of parameters given observed responses. Ordinarily, the above likelihood function is maximized by finding the parameter-value which makes the first derivative of the function equal to zero. This method is called Marginal Maximum-Likelihood estimation. Because calculating the integral implicit in the above function and its derivatives is cumbersome, this study will utilize a method based on Dempster, Laird, and Rubin's (1977) EM algorithm for actual parameter estimation.

Bock and Aitkin (1981) advanced the idea of using EM with probit-analysis inner cycles in the area of IRT parameter estimation, by replacing continuous theta with discrete theta points, chosen as convenient quadrature points for the integration. By following their approach, I will use discrete theta within IRT; of course Latent Classes are already discrete by definition.

EM Approach

The E-M algorithm of Dempster, Laird and Rubin (1977) makes it possible to compute maximum-likelihood estimates from incomplete data. As its name suggests, this algorithm is divided into two steps: the expectation step (E-step), and the maximization step (M-step). In this study the E-step entails estimating the proportion correct on each item and the proportion of subjects in the specified ability of the IRT group, as well as in given latent classes. For the IRT group, the M-step deals with maximizing a log-likelihood function to obtain new IRT item parameters, using estimated statistics from E-step. For the latent classes, the M-step consists in maximizing a log-likelihood function to obtain new latent-class model parameters. Cyclical application of E-step and M-step continues till a certain convergence accuracy is obtained.

E-step

After approximating the theta distribution, which may be continuous, by a discrete distribution over a finite number of points θ_q for the IRT group q, the posterior distribution over θ of the IRT group given a response vector $\underline{x}_j$ is the conditional probability of $\underline{x}_j$ given item parameters and ability divided by the marginal probability of $\underline{x}_j$ given item parameters, where the marginal probability of $\underline{x}_j$ is

$$P(\underline{x}|\underline{\zeta},\gamma-1) = \sum_q P(\underline{x}_j|\underline{\zeta},\underline{\theta}_q,\gamma-1)\cdot A(\theta_q)$$

Thus the posterior distribution of θ given $\underline{x}_j$ is

$$(4.4) \qquad P(\theta_q,|\zeta,\underline{x}_j,\gamma\text{-}1) = \frac{P(\underline{x}_j|\zeta,\theta_q,,\gamma\text{-}1)\cdot A(\theta_q)}{\sum_q P(\underline{x}_j|\zeta,\theta_q,\gamma\text{-}1)\cdot A(\theta_q)}$$

Notice that the finite set of points do not have to be ordered, and this equation can be expanded to include latent classes as well. Doing so,

$$(4.5) \qquad P(\theta_q,|\zeta,\underline{x}_j,\gamma\text{-}k) = \frac{P(\underline{x}_j|\zeta,\theta_q,,\gamma\text{-}k)\cdot A(\theta_q,)}{P(\underline{x}_j)}\cdot P(\gamma\text{-}k),$$

where $P(\underline{x}_j)$ is the marginal probability of $\underline{x}_j$, and is given by

$$(4.6) \qquad P(\underline{x}_j) = \left[\sum_q P(\underline{x}|\theta_q,\zeta,\gamma\text{--}1)\cdot A(\theta_q) \right]\cdot P(\gamma\text{-}1)$$

$$+ \sum_{k=2}^{K} P(\underline{x}_j|\zeta,\gamma\text{-}k)\cdot P(\gamma\text{-}k)$$

and

$$(4.7a) \qquad P(\underline{x}_j|\zeta,\theta_q,\gamma\text{-}1) = \prod_{i=1}^{I} P_{1i}(\theta_q)^{x_{ij}}\, Q_{1i}(\theta_q)^{1-x_{ij}}\cdot A(\theta_q)\cdot P(\gamma\text{-}1)$$

$$(4.7b) \qquad P(\underline{x}_j|\zeta,\gamma\text{-}k) = \prod_{i-1}^{I} P_{ki}^{x_{ij}} Q_{ki}^{1-x_{ij}}\cdot P(\gamma\text{-}k)$$

Also

$$(4.8) \qquad P(\gamma\text{-}1|\zeta,\underline{x}_j) = \frac{\sum_q P(\theta_q|\zeta,\underline{x}_j,\gamma\text{-}1)\cdot P(\gamma\text{-}1)}{P(\underline{x}_j)}$$

The posterior latent-class probability for class k is

$$(4.9) \qquad P(\gamma\text{-}k|\zeta,\underline{x}_j) = \frac{P(\underline{x}_j|\zeta,\gamma\text{-}k)\cdot P(\gamma\text{-}k)}{P(\underline{x}_j)}$$

Then the two sets of estimates needed to derive provisional estimates of propensity of correct response on item i given a point or a class are number of tries and number of correct responses. They are

$$(4.10a) \qquad \hat{R}_{liq} = \sum_{j=1}^{J} x_{ij}\cdot\frac{P(\underline{x}_j|\theta_q,\zeta,\gamma\text{-}1)\cdot A(\theta_q)}{P(\underline{x}_j)}\cdot P(\gamma\text{-}1)$$

$$(4.10b) \qquad \hat{N}_{liq} = \sum_{j=1}^{J} \frac{P(\underline{x}_j|\theta_q,\zeta,\gamma\text{-}1)\cdot A(\theta_q)}{P(\underline{x}_j)}\cdot P(\gamma\text{-}1)$$

$$(4.10c) \qquad \hat{R}_{ki} = \sum_{j=1}^{J} \frac{x_{ij}\cdot P(\underline{x}_j|\zeta,\gamma\text{-}k)}{P(\underline{x}_j)}\cdot P(\gamma\text{-}k)$$

$$(4.10d) \qquad \hat{N}_{ki} = \sum_{j=1}^{J} \frac{P(\underline{x}_j|\zeta,\gamma\text{-}k)}{P(\underline{x}_j)}\cdot P(\gamma\text{-}k)$$

M-step

The maximization step consists of maximizing likelihood equations entailed by the likelihood function (4.3). We first take the logarithm of the expression:

$$(4.11) \qquad \text{Log } L(\underline{\zeta}|\underline{X}) = \sum_{j=1}^{J} \ln \sum_{\gamma=1}^{K} P(\underline{x}_j|\underline{\zeta},\gamma)\cdot P(\gamma)$$

By differentiating (4.11) with respect to a generic parameter u, which can be either an item parameter of IRT or a latent class parameter, we get

$$(4.12) \qquad \frac{\partial \ln L(\underline{\zeta}|\underline{X})}{\partial u} = \sum_{j=1}^{J} \frac{\partial \ln \sum_{\gamma} P(\underline{x}_j|\underline{\zeta},\gamma)\cdot P(\gamma)}{\partial u}$$

$$= \sum_{j=1}^{J} \frac{1}{\sum_{\gamma} P(\underline{x}_j|\underline{\zeta},\gamma)\cdot P(\gamma)} \cdot \frac{\partial \sum_{\gamma} P(\underline{x}_j|\underline{\zeta},\gamma)\cdot P(\gamma)}{\partial u}$$

Notice that the first derivative of log likelihood with respect to IRT item parameters reduces to the ordinary MML problem for IRT as addressed by Bock and Aitkin (1981), as summands in the differential for $\gamma = 2,\ldots,K$ becomes zero. Likewise, the log likelihood differentiated with respect to latent-class parameters causes the IRT portion to drop out. Thus the M-step maximization for the IRT parameters and the latent class parameters can be carried out independently. Let us consider IRT item parameters first, then the latent class parameters.

IRT Item Parameters

Let u_i be a parameter of item i. By applying Bayes theorem, and Bock and Aitkin's idea of discrete theta points to approximate integration, the equation (4.12) can be rewritten as follows:

$$(4.13)\quad \frac{\partial \ln L(\underline{\zeta}|\underline{X})}{\partial u_i} = \sum_{j=1}^{J} \frac{1}{P(\underline{x}_j|\underline{\zeta})} \cdot \frac{\partial P(\underline{x}_j|\underline{\zeta},\gamma=1)}{\partial u_i} \cdot P(\gamma=1)$$

$$= \sum_{j=1}^{J} \frac{1}{P(\underline{x}_j|\underline{\zeta})} \cdot \frac{\partial \int_\theta P(\underline{x}_j|\theta,\underline{\zeta},\gamma=1) f(\theta) d\theta}{\partial u_i} \cdot P(\gamma=1)$$

$$= P(\gamma=1) \cdot \sum_{j=1}^{J} \int_\theta \frac{\partial P(\underline{x}_j|\theta,\underline{\zeta},\gamma=1)}{\partial u_i} \cdot \frac{f(\theta)}{P(\underline{x}_j|\underline{\zeta})}\, d\theta$$

$$= P(\gamma=1) \cdot \sum_{j=1}^{J} \int_\theta \frac{\partial P_i(\theta)}{\partial u_i} \cdot \frac{1}{P_i(\theta)Q_i(\theta)} \cdot$$

$$\left(\sum_{j=1}^{J} P(\theta|\underline{x}_j,\underline{\zeta},\gamma=1) \cdot (x_{ij} - P_i(\theta)) \right) d\theta$$

$$\doteq P(\gamma=1) \cdot \sum_{q} \frac{\partial P_i(\theta_q)}{\partial u_i} \cdot \frac{[R_{iq} - N_{iq} \cdot P_i(\theta_q)]}{P_i(\theta_q) \cdot Q_i(\theta_q)}$$

Specifically, for item parameters a_i and b_i,

$$(4.14a) \qquad \frac{\partial \ln L(\zeta|\underline{x})}{\partial a_i} = P(\gamma=1) \cdot \sum_q [R_{iq} - N_q \cdot P_i(\theta_q)] \cdot [-D(\theta_q - b_i)]$$

$$(4.14b) \qquad \frac{\partial \ln L(\zeta|\underline{x})}{\partial b_i} = P(\gamma=1) \cdot \sum_q [R_{iq} - N_q \cdot P_i(\theta_q)] \cdot D \cdot a_i$$

Marginal maximum likelihood estimates of item parameters are found as the roots of the expressions obtained by setting the derivatives of marginal log likelihood equal to zero. Item parameters can be obtained item by item within the M-step. The second derivatives which are necessary for a Newton-Raphson method of iterative parameter estimation are presented below.

$$(4.15a) \qquad \frac{\partial^2 \ln L}{\partial a_i^2} = -D^2 \cdot \sum_q (\theta_q - b_i)^2 N_{iq} P_i(\theta_q) Q_i(\theta_q)$$

$$(4.15b) \qquad \frac{\partial^2 \ln L}{\partial b_i^2} = -D^2 \cdot \sum_q a_i^2 N_{iq} P_i(\theta_q) Q_i(\theta_q)$$

$$(4.15c) \qquad \frac{\partial^2 \ln L}{\partial a_i \partial b_i} = D^2 \cdot \sum_q a_i (\theta_q - b_i)^2 N_{iq} P_i(\theta_q) Q_i(\theta_q)$$

The usual Newton-Raphson parameter estimation method can be described as an iterative application of the following equation till appropriate convergence criteria are met. The subscript "new" is for updated value and the subscript

"old" is for the value obtained from last iteration, or initial value if the iteration is the first one.

$$\begin{bmatrix} a_{i\ new} \\ b_{i\ new} \end{bmatrix} = \begin{bmatrix} a_{i\ old} \\ b_{i\ old} \end{bmatrix} - \begin{bmatrix} \frac{\partial^2 \ln L}{\partial a_i^2} & \frac{\partial^2 \ln L}{\partial a_i \partial b_i} \\ \frac{\partial^2 \ln L}{\partial b_i \partial a_i} & \frac{\partial^2 \ln L}{\partial b_i^2} \end{bmatrix}^{-1} \cdot \begin{bmatrix} \frac{\partial \ln L}{\partial a_i} \\ \frac{\partial \ln L}{\partial b_i} \end{bmatrix}$$

IRT Ability Parameters

Once estimates of item parameters are obtained, estimation of ability parameter can be carried out using several existing methods. One is the maximum likelihood method (MLE), which finds θ_j which maximizes the following likelihood expression (4.1). Note that the first subscripts of P and Q are because k=1, specifying the IRT group.

$$(4.1) \qquad L(\underline{x}_j|\theta_j,\underline{\zeta}) = \prod_{i=1}^{I} P_{1j}(\theta_j)^{x_{ij}} Q_{1j}(\theta_j)^{1-x_{ij}}$$

Differentiating (4.1) with respect to θ_j and dropping a subscript of k=1, the first derivative and second derivative are,

$$(4.16) \qquad \frac{\partial^2 \ln L(\underline{x}_j|\theta_j,\underline{\zeta})}{\partial\, \theta_j} = \sum_{i=1}^{I} D \cdot a_i \cdot (x_{ij} - P_i(\theta_j))$$

$$(4.17) \qquad \frac{\partial^2 \ln L(\underline{x}_j|\theta_j,\underline{\zeta})}{\partial\, \theta_j^2} = \sum_{i=1}^{I} D^2 \cdot a_i^2 \cdot P_i(\theta_j) \cdot Q_i(\theta_j)$$

Again utilize the Newton-Raphson method to estimate theta since there is no algebraic solution. The variance of this estimator is approximated by,

$$(4.18) \qquad \mathrm{Var}(\hat{\theta}_j) = \frac{1}{\sum_i D^2 \cdot a_i^2 \cdot P_i(\hat{\theta}) \cdot Q_i(\hat{\theta})}.$$

Other alternatives exist for estimating theta; they are Bayes modal estimator (maximum a posteriori, "MAP"), and Bayes expected a posteriori "EAP" estimator. Both are presented by Bock and Aitkin (1981). From Bayes' theorem the posterior conditional distribution of theta given $\underline{x} = \underline{x}_j$ is,

$$(4.19) \qquad f(\theta|\underline{x}_j,\underline{\zeta}) = \frac{P(\underline{x}_j|\theta,\underline{\zeta}) \cdot f(\theta)}{P(\underline{x}_j|\underline{\zeta})}$$

The MAP estimate of θ_j is the value which maximize the above function. The mode can be found by solving the following equation.

$$(4.20) \qquad \frac{\partial \ln f(\theta|\underline{x}_j,\underline{\zeta})}{\partial \theta_j} = \sum_{i=1}^{I} D \cdot a_i \cdot (x_{ij} - P_i(\theta_j)) + \frac{\partial f(\theta_j)}{\partial \theta_j} = 0$$

The second derivative is

$$(4.21) \qquad \sum_{i=1}^{I} D^2 \cdot a_i^2 \cdot P_i(\theta_j) \cdot Q_i(\theta_j) + \frac{\partial^2 f(\theta)}{\partial \theta_j^2}$$

The variance of MAP estimator of theta is approximated by

$$(4.22) \qquad \mathrm{Var}(\theta_j) = \frac{1}{\sum_{i=1} D^2 \cdot a_i^2 \cdot P_i(\hat{\theta}_j) \cdot Q_i(\hat{\theta}_j) + \frac{\partial^2 f(\hat{\theta})}{\partial \hat{\theta}_j^2}}$$

The EAP estimator for θ_j can be obtained by taking the expectation of function (4.19) through integrating over theta. In the E-step, integration was substituted by summation of values at fixed theta with appropriate weights. We can apply the same idea again in the M-step. Then the EAP estimator for θ_j is given by

$$(4.23) \qquad \hat{\theta}_j = E(\theta_j|\underline{x}_j) = \frac{\sum_q \theta_q \cdot P(\underline{x}_j|\theta_q) \cdot A(\theta_q)}{\sum_q P(\underline{x}_j|\theta_q) \cdot A(\theta_q)}$$

The variance of EAP estimator is approximately

$$(4.24) \qquad Var(\hat{\theta}_j) = \frac{\sum_q (\theta_q - \hat{\theta}_j)^2 \cdot P(\underline{x}_j|\theta_q) \cdot A(\theta_q)}{\sum_q P(\underline{x}_j|\theta_q) \cdot A(\theta_q)}$$

It is this author's preference to use the EAP estimator for the following reasons. The maximum likelihood estimator does not produce theta estimates for either all correct or all wrong response patterns. Also, unlikely patterns can cause the Newton-Raphson method to fail (Anderson, 1980). Since latent class responses are often unusual in the IRT sense, unless such response patterns are excluded before making theta estimates, theta estimation would be hampered. MAP is most compatible with a continuous density function of prior theta distribution. This is not always available. EAP does not have these problems possessed by MLE or MAP. Furthermore it can utilize the empirical discrete theta distribution generated in the E-step, obtained by summing the Nliq terms from (4.10b) over examinees, then normalizing. This is equivalent to Bock and Aitkin's Equation 15 (1981).

Latent Class Parameters

Let u_k be a parameter of latent class k. The likelihood equation (4.12) takes the form

$$(4.25)\qquad \frac{\partial \ln L(\zeta|\underline{x})}{\partial u_k} = \sum_{j=1}^{J} \frac{P(\underline{x}_j|\zeta,\gamma{=}k)\cdot P(\gamma{=}k)}{P(\underline{x}_j|\zeta)} \cdot \frac{\partial \ln P(\underline{x}_j|\zeta,\gamma{=}k)}{\partial u_k}$$

$$= \sum_{j=1}^{J} \frac{P(\underline{x}_j|\zeta,\gamma{=}k)\cdot P(\gamma{=}k)}{P(\underline{x}_j|\zeta)} \cdot$$

$$\sum_{i=1}^{I} \left(\frac{x_{ij}\,\partial \ln P_{ki}}{\partial u_k} + (1-x_{ij})^2 \cdot \frac{\partial \ln Q_{ki}}{\partial u_k} \right)$$

$$= \sum_{i}^{I} \sum_{j}^{J} \frac{P(\underline{x}_j|\zeta,\gamma{=}k)\cdot P(\gamma{=}k)}{P(\underline{x}_j|\zeta)} \cdot \frac{(x_{ij}-P_{ki})}{P_{ki}(1-P_{ki})} \frac{\partial P_{ki}}{\partial u_k}$$

$$= \sum_{i=1}^{I} \frac{R_{ki} - N_{ki}\cdot P_{ki}}{P_{ki}\cdot(1-P_{ki})} \frac{\partial P_{ki}}{\partial u_k}$$

Solving for a value of latent class parameter which makes the above equation zero may be done algebraically depending upon the type of latent classes. Let us consider two latent class types (L-type). Recall that a latent class k is characterized by a response vector $\underline{T}$ of I many 0's and 1's. The first L-type is characterized by one constant α_k representing the conditional probabilities of correct response to items for which $T_i = 1$, and another constant

β_k the probability of correct response to items for which $T_i = 0$. Then

$$\frac{\partial P_{ki}}{\partial \alpha_k} = 1 \qquad \text{for i such that } T_i = 1$$

$$\frac{\partial P_{ki}}{\partial \alpha_k} = 0 \qquad \text{for i such that } T_i = 0$$

also

$$\frac{\partial P_{ki}}{\partial \beta_k} = 0 \qquad \text{for i such that } T_i = 1$$

$$\frac{\partial P_{ki}}{\partial \beta_k} = 1 \qquad \text{for i such that } T_i = 0$$

Then the solutions of the following equations for α_k and β_k have closed expressions. The likelihood equations

$$\sum_{i}^{I} \frac{R_{ki} - N_{ki} \cdot P_{ki}}{P_{ki} \cdot (1-P_{ki})} \frac{\partial P_{ki}}{\partial u_k} = 0$$

imply that

(4.26a)
$$\hat{\alpha}_k = \frac{\sum_i R_{ki} \cdot T_{ki}}{\sum_i N_{ki} \cdot T_{ki}}$$

(4.26b)
$$\hat{\beta}_k = \frac{\sum_i R_{ki} \cdot (1-T_{ki})}{\sum_i N_{ki} \cdot (1-T_{ki})}$$

Let us consider the second L-type to be such that $\alpha_k = 1-\beta_k$. Then α_k can be estimated by

$$\alpha_k = \frac{\sum_i R_{ki} \cdot T_i + \sum_i (N_{ki}-R_{ki}) \cdot (1-T_i)}{\sum_i N_{ki}} \tag{4.27}$$

If an algebraic solution is not possible because of the nature of L-type then the Newton-Raphson method can be used.

Determining Convergence

It was considered that the estimation process was complete and convergence was obtained when any of updating parameter values do not differ from the previous corresponding values by more than a specified critical value. Since a is on a ratio scale and b is on an interval scale, the evaluation of the amount of updating value would reflect the nature of the two scales.

For a-parameters, the value δ_{ai} = (new a_i - old a_i)/old a_i is used to evaluate convergence, i.e. if the absolute value of δ_{ai} is less than or equal to a critical value, the new a_i is considered to be a converged value for item i. On the other hand, if the absolute value δ_{bi} = new b_i - old b_i is less than a critical value, the new b_i is accepted as a converged b value of item i.

PARAMETER ESTIMATES OF GENERATED DATA

Monte-Carlo data

The size of the simulated data set is 1000 simulees and 30 items. The proportional mixture of the IRT group and the Latent Class group was set to 70% and 30% respectively. They were simulated independently and later combined together and treated as one data set.

Seven hundred response vectors of the IRT group were generated using the 2-parameter logistic IRT model. Thirty pairs of item parameters a and b were simulated once and kept constant throughout the study; these are shown in Table 2. The a's are drawn from a normal distribution with a mean of 1.0 and standard deviation of 0.4, with only values greater than 0.3 kept. The b's are drawn from a normal distribution with a mean of 0.0 and standard deviation of 0.8. The relationship between a's and b's was not built into the design, so any correlation between a's and b's is by coincidental.

<insert Table 2 about here>

Seven hundred true trait values, θ, were generated from a standard normal distribution. The probability of a simulee j making a correct response on an item i, $P_i(\theta_j)$ was found using 2-parameter logistic function with the already simulated j, a_i, and b_i. Then for each (i,j) pair, a random number from a uniform distribution between 0.0 and 1.0 was generated. If the random number was less than $P_i(\theta_j)$, the simulee j correctly answered item i, and hence a score of 1 was given; otherwise a score of 0 was given. This procedure results in a 700x30 response matrix of simulees by items.

The parameters for the latent classes are conditional probabilities of correct responses on items given a latent class, and latent class probabilities, i.e. proportions of the several latent classes. Each latent class is specified by a particular response vector, the idealized response vector associated with that latent class. This study considered four latent classes, the response vectors of which appear in Table 3. The Monte-Carlo data sets are comprised of combinations of these four classes.

Five latent class types (L-type) were considered in different simulations. These L-types differing in the relationship of conditional

probabilities among items given a latent class. Let β_{i1k} be the conditional probability of correct response on an item i which has 1 for an idealized response of a given latent class k. Similarly, let α_{i0k} be the conditional probability of correct response on an item i which has 0 for an idealized response. For type one, β_{ik}'s are constant β; α_{ik}'s are also constant α and α is not equal to β. Type two is similar to type one except that α is equal to $(1-\beta)$. The L-type five is the freest of all three L-types; α_i's and β_i's are all allowed to be different to each other.

<insert table 3 about here>

The L-types three and four use information not only from idealized response vectors but also the conditional probabilities of IRT that are calculated by supposing an idealized response vector belongs to IRT group. These two types of latent classes are quite different from more conventional latent classes in the past, and they can express the biases of conditional probabilities often seen in the response distributions around highly unusual latent class in the "rule space". Bloxom (personal communication, 1988) pointed out that when the items are specified as correct, conditional probabilities are identical to those of 3-parameter IRT model. The first step in simulating data of type 3 and 4 is to estimate the theta of idealized response vector using the original a's and b's. Next is calculating the conditional probability for each item, a weighted sum of the idealized response and the $P_i(\theta)$ from IRT, with θ being an estimate for the idealized response vector.

In the case of L-type 3, let $\underline{T}_k$ be an idealized response vector of a latent class k, and θ_k be the estimated θ value from $\underline{T}_k$. Then the conditional

probability of correct response on item i given a subject in a latent class k is given by the following formula:

$$P_{ki} \begin{cases} = P_i(\theta_k) + \alpha_{k1} \cdot (1 - P_i(\theta_k)) & \text{when } T_{ki} = 1 \\ = \alpha_{k0} \cdot P_i(\theta_k) & \text{when } T_{ki} = 0 \end{cases}$$

Likewise conditional probabilities of L-type 4 can be given by the following formula.

$$P_{ki} \begin{cases} = P_i(\theta_k) + \alpha_{k1} \cdot (1 - P_i(\theta_k)) & \text{when } T_{ki} = 1 \\ = \alpha_{k0} & \text{when } T_{ki} = 0 \end{cases}$$

The constants used to combine the two are the parameters of L-types 3 and 4. Once we have a vector of conditional probabilities, the procedure for generating item responses after this point is identical to the one for the IRT group. All five L-types are presented in Table 4.

<insert Table 4 about here>

Results

The accuracy of the algorithm was examined with regard to the accuracy of estimated latent class parameters and IRT parameters which are standardized by making the estimated posterior theta distribution have a mean of 0.0 and a variance of 1.0. Distinct data sets were generated for each of the first four L-types. To challenge the procedure, multiple latent classes corresponding to the four patterns from Table 3 were included in each data set, since it is more difficult to estimate these parameters simultaneously, than if only one

type of aberration from IRT were present. For the estimation of parameters the input of latent classes (response patterns) was correct in each run, but solutions for all 5 L-type specifications were obtained. That is to say, for each data set, parameters were estimated five times with different L-type specifications, so that only one L-type was correct and the rest were erroneous. This was intended to test the robustness of estimated parameters to erroneous L-type specification. In order to find a possible systematic bias, 30 replications were carried out while keeping the configuration.

Experiment 1

This data set contained 700 IRT simulees, and 300 latent class simulees of four latent classes of L-type 1 and 2. The generating conditional probabilities of latent classes for these data are shown in Table 5. Accuracy of estimated IRT item parameters is summarized in Tables 6 and 7. Let $\underline{T}_k$ be a response pattern of a latent class k, and $\underline{T}_{ki}$ be a response on an item i. The first column represents the original generating values, and second through sixth columns represent the estimation results under the five different L-type specifications. The seventh column represents the result of parameter estimation by treating data as if it did not have any latent classes, but only an IRT group. This is fitting 2-parameter IRT model to the mixture data; without knowing anything about latent classes in the data, this is what one may attempt in practice in the first place. The eighth column represents the results of IRT parameter estimation using the data from only the 700 IRT simulees. Since the data do not contain any information about IRT apart from the 700 IRT simulees, any hybrid model parameter estimations cannot be more

accurate than those in this eighth column. The result of latent class parameter estimation is summarized in Table 8.

<insert tables 5, 6, 7, and 8 about here>

Correct L-type specification (L-type 1) was indistinguishable from other erroneous L-type specifications on correlations of estimated a-values with the original values. The bias was small (less than .06) for all L-types and very similar to the IRT-only (700) case. The ratio of the standard deviations showed that the a-values tend to be estimated slightly more variably than the original. The root mean squared deviations, RMSD's were quite small. In all respects, L-type specifications 1 and 5 produced the same results as IRT(700) case for the first 2 digits, which is the convergence criteria used for this study. The IRT(1000) case showed a large bias, small ratio of standard deviations, marginal correlations, and overall large error of estimation compared to any of L-type specification.

Estimated b-values showed their means to be greater than the original by .045 to .061 (similar to the values of .056 obtained using the IRT simulees only), and the error of the ratio of standard deviations to be less than .015 for all L-types. The correlation between the five estimated b-values and generating values showed that they are very close to the IRT-only case (all were 1.0). The RMSD were also very small. All L-type specifications, including erroneous ones, produced more uniformly accurately estimated item parameters than the No-Latent-class specification. In the run without latent classes specified [IRT(1000)] RMSD's were nearly 3 times as large for a-parameters and twice as large for b-parameters.

The latent class parameters were estimated very accurately, especially when the L-type is correctly specified. The four latent class proportions

were all within ± .007, and the conditional probabilities were within ± .007. L-type 2 specification calls for special attention because it is correct for three latent classes, but incorrect for one latent class #2. The accuracy of estimation has affected only the class which differs from the specified L-type. When L-type is correct the error of proportion estimates were within .007, conditional probabilities were within .001. Latent class #2 proportion estimate had error of .002, and probabilities of .150. It is surprising that even when conditional probabilities are estimated quite differently, the latent class proportion was estimated correctly. Apparently, erroneous L-type specifications affected only those latent classes which were different from the specified type.

For estimation runs that specified L-types 3, 4, and 5, latent class proportions were estimated within an error of .010. Estimated conditional probabilities were largest when L-type was specified as 3 and latent classes had less distinct conditional probabilities, such as .7 and .3, and .9 and .4. Again, accuracy of estimating latent class proportion was affected very little by L-type specification, but conditional probabilities of item response given latent class were very dependent upon L-type specification.

Experiment 2: Replication Study

The replication studies which was designed to see whether biases found on the first single experiment were persistent, and if so, to consider possible causes. Using the same item parameters and conditional probabilities described in the previous study, data sets were generated and estimated 30

times. The results are summarized in tables 9, 10 and 11.

<inset tables 9, 10 and 11 about here>

The mean estimation error of the a-parameter was less than .01 when L-type was specified as 1, 4, or 5. When items were separated into two groups (high and low) according to the values of original a values, mean estimation errors were larger for "high" items. But the error ratios were similar to each other and to overall items. Therefore no particular range of a values can be associated with greater estimation error. The overall accuracy of estimation RMSD's varied very little among all L-type specification. This indicates that the L-type specification differences in RMSD's of experiment 1 are due to the data variation.

The mean estimation error of the b-parameter was less than .001 when L-type was specified as 1 or 5. Other L-type specifications produced quite accurate estimates also, mean errors were all less than .012. In all respects, b-parameter estimation was very robust against erroneous L-type specifications.

Latent class parameters were estimated quite well especially when L-type specification is correct, i.e. L-type 1. When L-type was specified as 2 the accuracy of the latent class parameter estimates was unaffected by the errors of latent class 2 and showed very little sign of bias in both parameters, proportions and conditional probabilities. L-types 3 and 4 showed an expected amount of error for conditional probabilities just by the nature of latent class conditional probability constraints. However, the latent class proportions were generally estimated very well as seen from the maximum percentage error of 7% for the latent class 1 by L-type 3. L-type 5 produced very accurate results for both latent class parameters.

DISCUSSION

Although not all were reported in this paper, the estimated parameters of various Monte-Carlo data sets differing in the L-type, the number of latent classes and the conditional probabilities showed that the algorithm and the program were generally quite successful. In particular the item parameters of IRT and proportions of latent classes were fairly robust under erroneous specifications of L-types for the latent classes. Conditional probabilities of correct item response within latent classes were affected by the correctness of L-type specifications.

Table 12 shows the result of 30 replications on item parameter estimation of the IRT-only case. Since seeds to generate responses are identical to the one used for the replication study, the generated data are identical to the first 700 cases of the replication study data.

<insert table 12 about here>

The mean error of a's and b's of 30 converged replications showed that both parameters were estimated accurately without apparent biases within convergence criteria. This was equally true for both sets of items when items were separated into 2 sets according to their a-values where the one is items with a-values greater than 1.0 and the other with a-values less than 1.0. Since a is on a ratio scale it is most appropriate to examine the proportions; the mean proportion deviation's were .004 for both sets and the root mean square proportion deviation were .029 and .024 respectively. To compare the extent of errors of b-value estimation conditional on the original b-values, the items were separated into two sets, items with negative b-values were put into one set and items with positive b-values were put into the other. The

estimation bias of b-parameter was resulted to parameter values, that is to say the items with b-values greater than -.05 tend to be over estimated while items with b-values less than -.05 tend to be under estimated. This was also observed regardless of L-type specification and almost identical to IRT only(700) case. The presence of latent classes did not seem to affect estimated item parameters adversely any more than the estimation error itself, except when the L-type specification is very wrong such as L-type 3.

Undoubtedly the nature of the Bayesian estimation method itself would cause estimated parameters to regress toward the prior distribution, especially when the distribution of data is different from the prior, which was set to standard normal. If the error in the estimation is caused solely by the estimation method, then the variance of estimated item parameters would always be greater because the posterior theta distribution would be estimated to be less variant. The means should not be affected by the estimation method. The direction of biases as we see in the means of estimated item parameters in the IRT-only replication were the same for the replication studies with latent classes. In particular, a-values were over-estimated, even though the RMSD's were small. The characteristic of Bayesian estimation method, "regression toward mean", was evidenced in all experiments as we see in the ratios of standard deviations of estimated a-parameters to the standard deviations of original a-values. Since correlations among estimated values remain quite high, rescaling is the probable reason for the apparent bias.

It was found that the accuracy of conditional probabilities of item response within latent classes are sensitive to "correct" L-type specifications, (excluding cases such as L-type 1 being specified in estimation for L-type 2 data, since L-type 2 is a special case of L-type 1;

also excluded is the L-type 5 being specified in estimation, which places no constraints on conditional probabilities at all). When the data is L-type 3, any other systematic L-type would produce erroneous conditional probabilities except L-type 5. Proportions of latent classes were estimated quite accurately even when specified L-type were erroneous. While latent class proportions are estimated correctly, still the inaccuracy of the conditional probabilities is a troublesome problem, because in a real testing situation the correct L-type is not necessarily known. Future work should generate one general L-type that would subsume all systematic L-type cases discussed here (L-types 1, 2, 3, and 4). This can be obtained by having two separate regressions on conditional probabilities of IRT, one for the correct responses and the other for wrong responses of a latent class response pattern.

An experiment not reported in this paper examined a question concerning the effects of erroneous response patterns of latent classes, a wrong response pattern for a latent class was deliberately used to estimate model parameters. The "wrong" response pattern was wrong for 7 out of 30 items. The results showed that for all systematic L-types (1,2,3, and 4) specified estimation produced erroneous latent class parameters while IRT parameters were less affected adversely. But when L-type was specified as 5, i.e. no constraints on conditional probabilities across items, all of the model parameters were as accurate as when the response pattern of the latent class was correct.

In some cases of latent classes, it is very probable that one latent class type does not apply to all of the items; in such a case IRT might provide a convenient conditional probabilities for the rest of items. Future work should also address multiple constraints (L-types) on the items. For example, let us suppose that in a 20-items test L-type 1 applies to the first 10 items

and L-type 2 applies to the last 10. We might witness this case when a latent class is actually a composite of two latent classes differing in L-types. Also appropriate fit statistics of estimated parameters, and precise convergence criteria for both IRT item parameters and latent class parameters need to be investigated.

The hybrid model is applicable and especially able to reveal interesting cognitive structures when items are designed to diagnose critical performance in the measurement domain such as the items developed and data collected by Gitomer and Van Slyke (1988). They described erroneous trouble shooting skills in terms of the manifestation in the unique response patterns. The hybrid model is able to provide distribution of errors, conditional probabilities and subsequent posterior probability distributions. The model was also applied to the data with subjects using different solution strategies (Mislevy and Verhelst, 1989). A special case of two solution strategies, as described in their study, one set of subjects treat items seriously and answer with best of one's ability, and the other responds completely at random to the items. This type of data can be described by the hybrid model very well. The fit of the model with 1-parameter IRT and one latent class (fixed uniform conditional probabilities for all items) had nearly identical log-likelihood as the fit of three parameter IRT model with less degrees of freedom.

The potential utility of the hybrid model is especially great for items which are designed to diagnose relevant and critical performance rather than traditional ability test items. While traditional ability test items often have no clear relationship to a set of specific skills required to answer correctly on the item, cognitive model based test items have strong relationship between skills and knowledge used to answer the item.

In summary, the hybrid model provided unusual freedom which we had not experienced before, which comes from the fact that it is the union of continuous and discrete item response models. By varying group and state probabilities, the model can range from one extreme of IRT-only to the other extreme of latent-classes-only models, and also any combination in between. These latent classes do not have to be ordered at all, and hence they may be able to accommodate individual differences of the order of learning materials.

References

Andersen, E. B. (1977). Sufficient statistics and latent trait models. Psychometrika, 42, 69-81.

Angoff W. H. (1986, August). Validity: An evolving concept. Paper presented at ETS conference on validity, Princeton, NJ.

Bejar, I. I. (1980). A procedure of investigating the unidimensionality of achievement tests based on item parameter estimates, Journal of Educational Measurement, 17, 283-296.

Bejar, I. I. (1983). Introduction to Item Response Models and Their Assumptions. in Hambleton, R. K. ed. Applications of Item Response Theory, Educational Institute of British Columbia, Vancouver, B.C.

Birenbaum, M., & Tatsuoka, K. K. (1981). The effect of different instructional methods on error types and the underlying dimensionality of the test (Technical Report 81-3-ONR). Urbana, IL: University of Illinois, Computer-based Educational Research Laboratory.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord & Novick (Eds.), Statistical Theories of Mental Test Scores, Reading, MA.: Addison-Wesley.

Bock, R. D. (1967). Fitting a response model for dichotomous items. Paper read at the Psychometric Society Meeting, Madison, Wisconsin.

Bock, R. D., and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. Psychometrika, 46, 443-459.

Brown, J. S., and Burton, R. R. (1978). Diagnostic models for procedual bugs in basec mathematical skills. Cognitive Science, 2, 155-192.

Dayton, C. M., and Macready, G. B. (1976). A probabilitstic model for validation of behavioral hierarchies. Psychometrika, 41, 189-204.

Dempster, A. P., Laird, N. M., and Rubin D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society (Series B), 39, 1-38.

Formann, A. K. (1988) Latent class models for nonmonotone dichotomous items. Psychometrika, 53, 45-62.

Gitomer, D. H., and Van Slyke, D. H. (1988). Error analysis and tutor design. Machine-Mediated Learning, 2, 333-350.

Goodman, L. A. (1974). The analysis of qualitative variables when some of the variables are unobservable. Part I--A modified latent structure approach. American Journal of Sociology, 79, 1179-1259.

Heartel, E. H. (1984). Detection of a skill dichotomy using standardized achievement test items. Journal of Educational Measurement, 21, 59-72.

Holland, P. W. (1981). When are item response models consistent with observed data? Psychometrika, 46, 79-92.

Lazarsfeld, P. F.(1960). Latent structure analysis and test theory. In H. Gulliksen and S. Messick (Eds.), Psychological Scaling, Theory and Applications. New York: Wiley, 83-96.

Lazarsfeld, P. F., and Henry, N. W. (1968). Latent Structure Analysis. Boston: Houghton-Mifflin.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. Psychological Reports, 3, 635-694.

Lord, M. F. (1974). Estimation of latent ability and item parameters when there are omitted responses, Psychometrika, 39, 247-264.

Lord, M. F. (1980) Applications of Item Response Theory to Practical Testing Problems. NJ: Lawrence Erlbaum Associates.

Lord, M. F. & Norvick, M.R. (1968).Statistical Theories of Mental Test Scores. Reading. Mass.: Addison-Wesley.

Messick, S. (1980). Test validity and the ethics of assessment. American Psychologist, 1980, 35, 1012-1027.

Messick, S. (1984). The psychology of educational measurement. Journal of Educational Measurement, 21-3, 215-237.

McDonald, R. P. (1981). The dimensionality of tests and items, British Journal of Mathematical and Statistical Psychology, 34, 100-117.

Messick, S. (1980). Test validity and the ethics of assessment. American Psychologist, 35, 1012-1027.

Messick, S. (1984). The psychology of educational measurement. Journal of Educational Measurement, 21-3, 215-237.

Mislevy, R. J. (1983). Item response models for grouped data, Journal of Educational Statistics, 8, 271-288.

Mislevy, R. J. (1986). Bays modal estimation in item response models, Psychometrika, 51, 177-195.

Mislevy, R. J. and Verhelst, N. (in press). Modeling item response when different subjects wmploy different solution strategies, Psychometrika.

Paulson, A. J. (1982). A discrete latent-state approach to diagnostic testing, A research proposal submitted to the Office of Naval Research, Portland, OR: Portland State University.

Paulson, A. J. (1986a, May). Latent class representation of systematic patterns in test responses (Research Report). Portland, OR: Portland State University.

Paulson, A. J. (1986b, June). Estimation of parameters in latent class models with constraints on the parameters (Research Report). Portland, OR: Portland State University.

Reckase, M. D. (1982). Multidimensional IRT models. A paper presented at 1982 CAT conference at Minneapolis.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. Applied Psychological Measurement, 9, 401-412.

Riley, M. S. and Greeno, J. M. (1988). Developmental analysis of understanding language about quantities and of solving problems. Cogniton and Instruction, 5(1), 49-101.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometric Monograph, 17.

Samejima, F. (1972). A general model for free-response data. Psychometric Monograph Supplement, 18.

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. Psychometrika, 39, 111-121.

Siegler, R. S. (1981). Developmental sequences within and between concepts. Monographs of The Society for Research in Child Development, 46.

Stout, W. F. (1984). The statistical assessment of latent trait dimensionality in psychological testing, (final Report 8-82-ONR), University of Illinois, Department of Mathematics, Urbana-Champaign.

Stout, W. F. (1986). Issues concerning local independence, dimensionality, and bias: advances in testing for unidimensionality and bias with a large-scale simulation study. A paper presented at ONR contractors meeting, Gatlinberg, TN.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. Psychometrika, 52, 589-617.

Sympson, J. B. (1977). A model for testing with multidimensional items. Pater presented at the Computerized Adaptive Testing Conference, Minneapolis, MN.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. Journal of Educational Measurement, 20, 345-354.

Tatsuoka, K. K. (Ed.) (1984). Analysis of errors in fraction addition and subtraction problems (Research Report). Urbana, Ill.: University of Illinois, Computer-based Educational Research Laboratory.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. Journal of Educational Statistics, 12, 55-73.

Tatsuoka, K. K., & Tatsuoka, M. M. (1982). Detection of aberrant response patterns. Journal of Educational Statistics, 7, 81-96.

Tatsuoka, K. K., & Tatsuoka, M. M. (1987). Bug distribution and statistical pattern classification, Psychometrika, 2, 193-206.

Traub, R. E. (1983). Apriori considerations in choosing an item response model. In Hambleton, R. K. (ed.) Applications of Item Response Theory. Educational Institute of British Columbia, Vancouver, B. C.

Wang, M. M. (1986). Fitting a unidimensional model to multidimensional item response data. Paper presented at ONR contractors meeting at Gatlinberg.

Wang, M. M. (1987). Measurement bias in the application of a unidimensional item response data. Paper presented at ONR contractors meeting, Columbia, SC.

Wilcox, R. R. (1979). Achievement tests and latent structure models. British Journal of Mathematical and Statistical Psychology, 32, 61-71.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97-116.

Yamamoto, K. (1983). Derivation and test of predictions of a discrete latent state model for signed number addition test performance. Unpublished master's thesis, Portland, OR: Portland State University.

Yamamoto, K. and Tatsuoka, K. (1985, March). CAT Misconception Analysis: Modifications and Conclusions. Paper presented at 1985 AERA Annual Meeting; Chicago, IL.

Table 1

A Model That Specify Conditional Probabilities and State Probabilities for Absolute Value Component of Signed-number Addition Problems

			Item types				
K	State of subject	P(k)	L+-S	-L+S	-L+-S	-S+L	S+-L
1	correct rule	p(k=1)	$1-\varepsilon_1$	$1-\varepsilon_1$	$1-\varepsilon_1$	$1-\varepsilon_1$	$1-\varepsilon_1$
2	always adds	p(k=2)	ε_2	ε_2	$1-\varepsilon_2$	ε_2	ε_2
3	always subtract	p(k=3)	$1-\varepsilon_3$	$1-\varepsilon_3$	$\varepsilon 3$	$1-\varepsilon_3$	$1-\varepsilon_3$
4	always guessing	p(k=4)	$1-\varepsilon_{41}$	$1-\varepsilon_{42}$	$1-\varepsilon_{43}$	$1-\varepsilon_{44}$	$1-\varepsilon_{45}$

Table 2

Original Item Parameter Values and Mean and Standard Deviations of Theta Values for Monte Carlo Data

Item	A	B	Theta
1	1.4895	-.1883	
2	1.1529	-.3417	
3	1.1676	.1757	
4	1.0199	-.8639	
5	.7947	-.0959	
6	.6503	-.2416	
7	.4072	-1.3513	
8	1.3647	-1.1174	
9	1.5884	1.2061	
10	0.4290	-.3831	
11	1.8876	-.9237	
12	0.7353	1.2128	
13	0.5179	1.3195	
14	1.1703	-.1527	
15	0.6759	.0207	
16	1.0176	1.8858	
17	1.6346	1.0201	
18	1.5207	-.9268	
19	0.5292	.7802	
20	1.0727	.7637	
21	1.5662	.3522	
22	0.8882	-.0143	
23	1.0497	-1.5994	
24	1.5700	.5674	
25	0.8052	-.2334	
26	0.5834	.1844	
27	1.2640	.1821	
28	1.4731	-.2357	
29	1.4369	.1919	
30	1.0620	-1.1936	
Mean	1.0842	0.0000	-.0130
Sd	.4036	.8377	1.0264

Table 3

Response patterns of Four Latent Classes

Latent class	Idealized response pattern
1	11111111111111100000000000000
2	00000000000000011111111111111
3	11111000001111100000111110000
4	00000111110000011111000001111

Table 4

Parameters of Conditional Probabilities for

Five Latent Class Types (L-type)

	Sample response pattern (T_{ki})	
L-type	1	0
1	α_k	β_k
2	α_k	$1-\alpha_k$
3	$P_i(\theta_k+\alpha_{k1}(1-P_i(\theta_k))$	$\alpha_{k0}P_i(\theta_k)$
4	$P_i(\theta_k+\alpha_{k1}(1-P_i(\theta_k))$	α_{k0}
5	α_{ki}	$1-\alpha_{ki}$

Table 5

Configuration of Data Set

IRT group

Number of subjects 700

Latent class group

1. Number of subjects 120
   L-type 2
   Response pattern (T_i) 11111 11111 11111 00000 00000 00000
   Conditional probabilities 0.7 for item i if $T_i = 1$
   0.3 for item i it $T_i = 0$

2. Number of subjects 90
   L-type 1
   Response pattern (T_i) 00000 00000 00000 11111 11111 11111
   Conditional probabilities 0.9 for item i if $T_i = 1$
   0.4 for item i it $T_i = 0$

3. Number of subjects 60
   L-type 2
   Response pattern (T_i) 11111 00000 11111 00000 11111 00000
   Conditional probabilities 0.9 for item i if $T_i = 1$
   0.1 for item i it $T_i = 0$

4. Number of subjects 30
   L-type 2
   Response pattern (T_i) 00000 11111 00000 11111 00000 11111
   Conditional probabilities 0.9 for item i if $T_i = 1$
   0.1 for item i it $T_i = 0$

Table 6

Comparison of Estimated A-Parameters by All Five L-types, IRT-only(1000), IRT-only(700), and Original A-Parameters

		L-types						
	Original	1	2	3	4	5	IRT(1000)	IRT(700)
Means	1.084	1.139	1.130	1.098	1.132	1.141	.737	1.142
SDa/SDa	1.000	1.192	1.159	1.102	1.182	1.197	.747	1.194
Correlation								
Original	1.000	.947	.947	.952	.946	.946	.710	.949
1		1.000	.998	.993	.999	1.000	.678	.999
2			1.000	.991	.998	.998	.700	.997
3				1.000	.993	.992	.725	.995
4					1.000	.999	.687	.999
5						1.000	.675	.999
IRT only(1000)							1.000	.684
IRT only(700)								1.000
Regression constants								
Intercept		.027	.023	.019	.027	.028	.055	.027
Slope		.928	.938	.970	.934	.926	1.397	.926
SSD								
Original	.000	.623	.597	.522	.632	.639	2.966	.600
1		.000	.024	.109	.008	.001	4.183	.011
2			.000	.121	.025	.026	3.746	.037
L-type 3				.000	.088	.108	3.166	.072
4					.000	.009	4.001	.010
5						.000	4.248	.012
IRT only(1000)							.000	1.684
IRT only(700)								.000
RMSD	.000	.171	.162	.138	.168	.175	.449	.171

Table 7

Comparison of Estimated B-Parameters by All Five L-types, IRT-only(1000), IRT-only(700), and Original B-Parameters

		L-types						
	Original	1	2	3	4	5	IRT(1000)	IRT(700)
Means	.000	.057	.045	.053	.061	.057	-.090	.056
SDb/SDb	1.000	.994	.987	.999	.995	.994	.929	.990
Correlation								
Original	1.000	.988	.989	.987	.988	.988	.979	.988
1		1.000	1.000	1.000	1.000	1.000	.986	1.000
2			1.000	1.000	1.000	1.000	.987	1.000
3				1.000	1.000	1.000	.986	1.000
4					1.000	1.000	.986	1.000
5						1.000	.986	1.000
IRT only(1000)							1.000	.986
IRT only(700)								1.000
Regression constants								
Intercept		-.056	-.045	-.052	-.060	-.056	.094	-.056
Slope		.990	.999	.984	.987	.989	1.040	.993
SSD								
Original	.000	.500	.452	.532	.514	.499	.881	.509
1		.000	.004	.010	.001	.000	.589	.002
2			.000	.018	.007	.004	.534	.005
L-type 3				.000	.009	.011	.578	.015
4					.000	.001	.607	.003
5						.000	.594	.002
IRT only(1000)							.000	.517
IRT only(700)								.000
RMSD	.000	.141	.131	.144	.145	.141	.198	.142

Table 8

Comparison of Estimated Latent Class Parameters by All Five L-types Against Original Values

		Mean error of conditional probabilities	MSD	P(k)	SD
L-type = 1					
L-class	1	-.007	.000	.127	.000
L-class	2	-.002	.000	.091	.000
L-class	3	.007	.000	.029	.001
L-class	4	-.001	.000	.060	.001
L-type = 2					
L-class	1	.000	.000	.127	.000
L-class	2	-.150	.023	.088	.000
L-class	3	.000	.000	.029	.001
L-class	4	.000	.000	.059	.001
L-type = 3					
L-class	1	-.034	.028	.130	.000
L-class	2	-.014	.012	.092	.000
L-class	3	.005	.004	.030	.001
L-class	4	-.000	.002	.060	.001
L-type = 4					
L-class	1	.005	.007	.130	.000
L-class	2	.002	.002	.090	.000
L-class	3	.008	.002	.029	.001
L-class	4	.001	.001	.060	.001
L-type = 5					
L-class	1	-.006	.002	.127	.000
L-class	2	-.002	.001	.091	.000
L-class	3	.006	.004	.030	.001
L-class	4	-.001	.001	.060	.001

Table 9

Comparison of Estimated A-Parameters by All Five L-types

Over 30 Replications Against the Original

Level	Item	$\bar{R}_{a,\hat{a}}$	Mean $\hat{a}$	$\overline{a-\hat{a}}$	$\overline{a-\hat{a}}$ / a	$\overline{SD}_{\hat{a}}$	$SD_{\hat{a}}/SD_a$	$\overline{RMSD}_{\hat{a}}$
L-type 1								
Low	15	.890	.740	.004	.007	.250	1.141	.123
High	15	.760	1.413	.011	.008	.286	1.325	.196
All	30	.930	1.077	.008	.008	.433	1.074	.165
L-type 2								
Low	15	.889	.737	.007	.011	.248	1.133	.122
High	15	.753	1.406	.018	.013	.286	1.324	.198
All	30	.928	1.072	.012	.012	.431	1.068	.166
L-type 3								
Low	15	.878	.719	.026	.028	.221	1.006	.119
High	15	.712	1.375	.049	.033	.258	1.193	.201
All	30	.926	1.047	.037	.030	.409	1.013	.167
L-type 4								
Low	15	.889	.744	-.000	.003	.257	1.171	.127
High	15	.753	1.414	.010	.008	.283	1.309	.195
All	30	.928	1.079	.005	.005	.433	1.072	.166
L-type 5								
Low	15	.889	.741	.004	.007	.251	1.143	.123
High	15	.758	1.414	.010	.008	.287	1.328	.197
All	30	.929	1.077	.007	.007	.434	1.076	.166

Table 10

Comparison of Estimated B-Parameters by All Five L-types

Over 30 Replications Against the Original

Level	Item	$\bar{R}_{b,\hat{b}}$	Mean $\hat{b}$	$\overline{b-\hat{b}}$	$\overline{SD}_{\hat{b}}$	$SD_{\hat{b}}/SD_b$	$RMSD_{\hat{b}}$
L-type 1							
Low	15	.962	-.675	.019	.528	1.085	.150
High	15	.973	.676	-.019	.587	1.063	.148
All	30	.986	.000	-.000	.877	1.047	.151
L-type 2							
Low	15	.962	-.683	.027	.533	1.096	.153
High	15	.973	.661	-.005	.587	1.045	.145
All	30	.986	-.011	.011	.873	1.042	.151
L-type 3							
Low	15	.964	-.673	.017	.536	1.102	.152
High	15	.973	.664	-.008	.590	1.069	.149
All	30	.986	-.004	.004	.876	1.045	.152
L-type 4							
Low	15	.962	-.671	.014	.525	1.078	.149
High	15	.973	.676	-.020	.588	1.064	.149
All	30	.986	.003	-.003	.875	1.044	.151
L-type 5							
Low	15	.962	-.675	.019	.528	1.085	.150
High	15	.973	.675	-.019	.587	1.063	.147
All	30	.986	.000	-.000	.877	1.047	.151

Table 11

Comparison of Estimated Latent Class Parameters by All Five L-types over 30 Replications Against the Original Values

	Mean error of conditional probabilities	MSD	Mean P(k)	MSD
L-type 1				
L-class 1	.001	.000	.120	.000
L-class 2	.001	.000	.091	.000
L-class 3	-.000	.000	.060	.000
L-class 4	.000	.000	.030	.000
L-type 2				
L-class 1	.000	.000	.119	.000
L-class 2	-.150	.023	.087	.000
L-class 3	.000	.000	.060	.000
L-class 4	.000	.000	.030	.000
L-type 3				
L-class 1	-.019	.027	.127	.000
L-class 2	-.015	.014	.091	.000
L-class 3	-.000	.003	.061	.000
L-class 4	.002	.003	.031	.000
L-type 4				
L-class 1	.013	.009	.124	.000
L-class 2	.003	.002	.091	.000
L-class 3	.001	.001	.061	.000
L-class 4	.002	.002	.030	.000
L-type 5				
L-class 1	.001	.002	.120	.000
L-class 2	.001	.002	.091	.000
L-class 3	.000	.002	.060	.000
L-class 4	-.000	.003	.030	.000

Table 12

Comparison of Estimated Item Parameters over 30 Replications

700 IRT Simulees Against the Original Values

A - Parameter

Level	Item	$\bar{R}_{a,\hat{a}}$	Mean $\hat{a}$	$\overline{\hat{a}\text{-}a}$	$\overline{\hat{a}\text{-}a/a}$	$\overline{SD}_{\hat{a}}$	$\overline{SD_{\hat{a}}/SD_a}$	$\overline{RMSDa}$
Low	15	.893	.742	.002	.004	.250	1.141	.121
High	15	.769	1.332	.005	.004	.288	1.332	.194
All	15	.932	1.081	.003	.004	.435	1.079	.164

B - Parameter

Level	Item	$\bar{R}_{b,\hat{b}}$	Mean $\hat{b}$	$\overline{\hat{b}\text{-}b}$	$\overline{SD}_{\hat{b}}$	$\overline{SD_{\hat{b}}/SD_b}$	$\overline{RMSDb}$
Low	15	.963	-.670	.014	.526	1.081	.148
High	15	.973	.673	-.017	.582	1.055	.145
All	15	.986	.001	-.001	.872	1.041	.149